International Journal of Population Data Science

Journal Website: www.ijpds.org

Managing Emerging Data Types in the NGLMS

Farrow, James^{1*}

¹Farrow Norris / SANT DataLink

Objectives

We describe the management system used by the Next Generation Linkage Management System (NGLMS) built for SA.NT DataLink in Adelaide, Australia. The NGLMS is a bespoke system built on freely available open source components where a graph (in the computer science sense) structure is used to store a 'more natural' representation of linked records explicitly in a graph database: records as vertices and relationships as edges between vertices.

Approach

The NGLMS is designed to manage linked data effectively and permit fast individual cluster extraction while retaining rich relationship information. It holds probabilistic and statically-linked data by storing all significant pair-wise relationships between records as edges in a graph, allowing clustering with different parameters to be performed dynamically. Records are heterogeneous and may contain different data types: birth records, hospital separations, census data, pharmaceutical prescriptions, educational data. The relationships between records are also heterogeneous and may represent arbitrary relationships not just a probabilistic record similarity. For example, familial (parent/child), tribal kinship structures, genomic (and other omic) information, employer/employee relationships, educational information, living arrangements, census information, and so on. Storing this information allows for richer queries than just 'do these records represent the same entity'. For example a single rich query to the database could be 'find all records of all siblings', 'create genealogies based on birth information', 'create household groups based on census/cohabitation information', or 'find employees working in areas affected by recent floods with hospitalisations during that time period.'

Results

We present details of the loading of birth and perinatal data incorporating parent (mother and father) relationships for some South Australian datasets and the technical configuration of the NGLMS to support this. We discuss the queries made possible as a result. Rich non-traditional data is stored in the same manner as probabilistic record similarities and has allowed clustering queries which mix explicit deterministic statements about the data and probabilistic statements concerning record relationships.

Conclusion

Rich queries over data may be expressed by storing rich heterogeneous information about records and relationships explicitly as a graph and by determining clusters late in the extraction process. Modern graph database technologies make this effective even in the face of datasets containing 10's to 100's of million records and billions of edge relationships.

*Corresponding Author:

Email Address: james@fn.com.au (J. Farrow)